DOCUMENT RESUME

ED 220 523 TM 820 543

AUTHOR McLean, Les
TITLE Contemporary Approaches to Unit of Analysis and Site Variability Issues from the Follow Through Evaluation.
PUB DATE Mar 82
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association, (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Classroom Environment; *Data Analysis; Elementary Secondary Education; Multiple Regression Analysis; *Predictor Variables; Sampling; *Site Analysis; *Validity
IDENTIFIERS Evaluation Problems; *Project Follow Through; Unit of Analysis Problems

ABSTRACT

The issues and some proposed solutions regarding Follow Through (FT) site variability are examined with a review of developments in FT evaluation. The role of adjusted site means with differences within sponsors and between sponsors and background characteristics is discussed to determine whether adjusted means are the preferred measures of model effectiveness. In a Big City Group, attrition bias in data for non-FT and FT site analysis is considered. Improvements in measurement are shown in the sampling of content and behavior, including the use of computer systems with broad content samples. These procedures can eliminate reliance on multiple choice questions and the use of classroom process data with student reports on opportunity to learn (OTL) data expanding the dimensions of variability. A comtemporary model which crosses class type with school sites illustrates the multilevel regression analysis. Student scores are the dependent variable; and class type, sex, OTL class mean and individual math ability are the independent variables. The significant role of OTL to the stepwise fitting of the model is shown. (CM)

Contemporary Approaches to Unit of Analysis and Site Variability

Issues from the Follow Through Evaluation[1]

Les McLean

Head, Educational Evaluation Centre

The Ontario Institute for Studies in Education

Looking back at the Follow Through evaluation from a 1982 perspective, one is struck by just how much things have changed in ten years. There is much more respect for contextual effects and the need to link achievement measures more closely to curriculum content. Contemporary approaches, therefore, are not just fancier regression analyses but include more complex designs, more exploration of the data and fancier regression analyses. In this paper, some Follow Through evaluation data will be revisited briefly, but one or two contemporary examples will serve better to illustrate contemporary approaches. The purpose is to discuss the issues and some proposed solutions rather than to argue the Follow Through site variability question one way or the other.

Follow Through Revisited

Exploration. Raw score site means are displayed in Figure 1 for several outcome variables from the FT evaluation. These are continued in Figure 2, where background variables are added (again, site means). These figures provide a graphic display of the within site, across

---

[1] A paper presented as part of the symposium "The Site Variability Issue in Follow Through Revisited: Some New Data, Some New Methodologies and New Insights." AERA annual meeting, New York, N.Y., March 19-23, 1982.

sponsor distributions. Examined in this form, it is clear that, as the Abt Associates evaluation said (Stebbins et al., 1977), "the effectiveness of each FT model varied substantially from site group to site group." It is also plausible, though not as consistently clear, that "overall model averages varied little in comparison." Model differences appear to be larger on math computations than reading, but the range of site means is large in both cases.

What is also clear is that the sites within each sponsor varied substantially on background characteristics (Ethnic-linguistic, SES, WRAT) as well as in "effectiveness" (scores on the MAT subtests). It does not seem, however, that the average and range of site background characteristics differed greatly from sponsor to sponsor.

Looking just at the Follow Through groups, the smallest range of means on reading is 6 points (Behav. Anal.), a grade equivalent range of about 11 months. The largest (Resp. Educ.) is 10 points. Sponsor means range from 16 to 18. The exploration has yet to shake the plausibility of the Abt finding on site variability.

Confirmation. Bereiter and Kurland (1978 and in press) took a sensible tack (straightforward and conventional, Bereiter and Kurland, 1978, p. 3) and adjusted site means for background characteristics. Insofar as achievement is related to SES and the like, some of the variance we see can be attributed to background. Ethnic-linguistic and SES measures are correlated with achievement, so one expects covariance analysis to affect the results--and it does. Differences among sponsor means that were not statistically significant before become significant.

A previous exploratory observation was confirmed by the covariance analysis when the differences among adjusted means were seen to be virtually identical to those among unadjusted means. Background differences were similar from sponsor to sponsor. The covariance adjustment (shown by Bereiter and Kurland to be robust) has reduced the error variance and allowed us to infer with some confidence that the differences we were observing between and among sponsor means are not likely sampling fluctuations or other statistical artifacts.

But what of site variability? Estimates of between-sponsor differences were unchanged, but the estimates of within-sponsor variability (site variability) were reduced. Now, overall (adjusted) model averages vary more in comparison to the variation in adjusted means from site to site. What remains to ask is whether the adjusted means are the preferred measures of model "effectiveness."

That this may not be completely straightforward was argued by Cronbach, Rogosa, Floden and Price (1977), and it doesn't seem completely straightforward to take the reduced variance estimate as proof that differences among models previously regarded as modest in context should now be regarded as important. No doubt Kurland will clarify the matter in his paper (Kurland, 1982). Before considering other confirmatory analyses, consider one more contextual issue raised by exploratory analysis.

The Big City Group. Substantial attrition did occur over the three years of the evaluation (Stebbins et al., 1977, p. 82), but AAI were persuaded that no bias resulted. Pursuing the attrition matter, McLean (1978) plotted differential attrition (FT vs NFT) against

differential WRAT scores, with a result (Figure 3) that suggested an attrition bias acted against the non-Follow Through groups. (The trend from upper left to lower right is significant: $r = -.51$, $p < .01$.) More low-scoring students dropped out of the Follow Through than the non-Follow Through groups.

The bottom right-hand quadrant in Figure 2 contains sites for which the FT attrition exceeded the NFT and for which the FT WRAT scores were lower than NFT. When these sites were identified, all but two of the big city sites were there, and the other two were nearby. The sites in the upper left-hand quadrant turned out to be smaller communities, suggesting a contextual effect that had not been turned up in the omnibus analyses. This type of analysis has been followed up by Gersten (1982).

A purely site-level analysis cannot be refined to any extent (by grouping, for example) because the sample size is too small. Combining student-level and site-level data would be an attractive alternative, to be discussed in the last section. First, however, consider how content and measuring techniques might affect site and model variability.

Improvements in Measurement and in the Sampling of Content and Behavior

The narrow coverage of early childhood outcomes was criticized by House et al., (1978) and a number of sponsors felt keenly that the measures selected for the evaluation were not valid indicators of the effectiveness of their programs. Certainly the multiple-choice format dictated by the technical and financial constraints placed on the evaluation severely restricted the sample of student behavior obtained from these nine-year-olds (not to speak of the five- and six-year-olds).

In short, the observed variability was a drastically reduced sample of reality.

Since large item pools are available that may be used with item sampling techniques, there is no longer any excuse for poor curriculum coverage in a large, important study. With a total cost estimated at $30-$50 million (House et al., 1978, p. 129; 1977 dollars), the Follow Through evaluation certainly qualified as large and important.

Modern computer systems have also removed the need to rely exclusively on multiple-choice questions in large evaluations or assessments. As an example, the 1981 Field Trials of the Ontario Assessment Instrument Pools in mathematics and English involved over 37,000 students in grades 7 to 10 in 180 schools, as well as 1000 English and 600 mathematics instruments, most of which required a constructed response.

All responses were entered to computer files, checked and readied for scoring in eight weeks, by specially trained clerks using custom computer programs. Subsequent analysis steps were largely the same as those that confronted the AAI staff, with two important elaborations. First, the content sample was broader and more finely stratified. The mathematics content included 55 terminal objectives, for example, each of which was represented in the field trial by six examples. Sixteen topics (analogous to subtests) were chosen for summaries (e.g., whole numbers, decimals, fractions, integers, algebra and the like--elementary and intermediate).

The second elaboration was the inclusion of classroom process data, along with student reports on opportunity-to-learn (OTL). These latter were suggested by association with the Second International

Mathematics Study (SIMS), now almost complete in 23 countries. In SIMS, both student and teacher OTL reports were collected, along with elaborate reports on teacher math constructs and classroom procedures. An important practical result of these elaborations is that the dimensions of variability expand exponentially, demanding new data analytic approaches. There are many from which to choose, and this paper might better have been entitled "A few of the simpler contemporary[2] approaches . . . "

Contemporary Example

Three Class-types Crossed With Nine School Nested Within Four School Boards

Board 1	N	Board 2	N	Board 3	N	Board 4	N
S1	Class 1 (13) ------- Class 3 (27)	S4	Class 1 (16) Class 2 (31) -------	S6	Class 1 (15) Class 2 (29) -------	S8	Class 1 (11) ------- Class 3 (14)
S2	Class 1 (13) Class 2 (29) -------	S5	Class 1 (14) ------- Class 3 (23)	S7	Class 1 (10) ------- Class 3 (25)	S9	------- Class 2 (22) Class 3 (29)
S3	Class 1 (11) ------- Class 3 (25)		-------		-------		-------

Class 1: Basic level, low achievement
Class 2: General level, cross section
Class 3: Advanced level, high achievement

[2] contemporary . . . Belonging to same time or of same age, esp. as oneself; (ultra) modern in style or design (Oxford Pocket Dictionary, 6th ed., 1978).

Variance Decomposition

Boards	112.7	1%
Schools/Bd	1394.2	13%
Class Level	2867.5	26%
Class-L x School	390.2	4%
Within School	6062.7	56%
Total	10827.3	
Total N	357	

Multilevel regression. Burstein[3] advocates fitting regression models containing both aggregated and student-level data. Such a model for the data in the contemporary example might include:

Dependent Variable: (student level)	Total math score
Independent Variables:	
1. Class type (categorical)	Basic, General, Advanced
2. Sex (categorical)	Female, Male
3. Opportunity to Learn (class mean)	Scale: 0 to 20
4. Relative Math Ability (student level)	Total Math - Class Mean Subtest-- Prerequisites

[3]Burstein, Leigh. Explanatory models using between and within class regression: basic concepts and an example. Paper presented at the data analysis workshop, Second International Mathematics Study, Toronto, Canada, December 7-11, 1981.

The result of fitting such a model (stepwise)

Var	R^2	Inc. in R^2	Reg. Coeff.	Simple Correlation	Partial Correlation
OTL	0.09	0.09	.42	.29	.18
Class	0.13	.04	1.13	.28	.17
Sex	.15	.02	-1.47	-.16	-.14
Rel. Math	.16	.01	0.06	.27	.13

Figure 4 is a scatterplot of OTL with math score (class means). It is interesting to observe that OTL is a powerful variable (pooled within class correlation with score is 0.5) over and above differences among classes and schools.

The lesson this author draws with regard to Follow Through is that the issue of site variability probably cannot be adequately explored with the data as collected. We might best move on to other tasks.

References

Bereiter, Carl and Kurland, Midian. Were some Follow Through models more effective than others? AERA annual meeting, Toronto, 1978.

Bock, G., Stebbins, L. B. and Proper, E. C. Education As Experimentation: A Planned Variation Model. Vol. IV. B. Effects of Follow Through Models. Cambridge, Mass.: Abt Associates, Inc., 1977.

Burstein, Leigh. The analysis of multilevel data in educational research in evaluation. In David Berliner (Ed.), Review of Research in Education, Vol. 8. Washington, D.C.: AERA, 1980, pages 158-233.

Cronbach, L. J. Research on classroom and schools: formulation of questions, designs and analysis. Occasional paper: Stanford Evaluation Consortium, 1976.

Cronbach, L. J., Rogosa, D. R., Floden, R. E. and Price, G. G. Analysis of covariance in nonrandomized experiments: parameters affecting bias. Occasional paper: Stanford Evaluation Consortium, 1977.

Gersten, Russell M. A new look at site variability and durability of effects in Follow Through: some new empirical findings. AERA annual meeting, 1982.

House, Ernest R., Glass, Gene V., McLean, Leslie D. and Walker, Decker F. No simple Answer: Critique of the Follow Through Evaluation. Harvard Educational Review, 1978, 48(2), 128-160.

Keesling, J. Ward and Wiley, David E. Regression models for hierarchical data. Psychometric Society, 1974.

Kurland, D. Midian. Methodological and policy implications of the Follow Through results: matching the statistical model to the research question. AERA annual meeting, 1982.

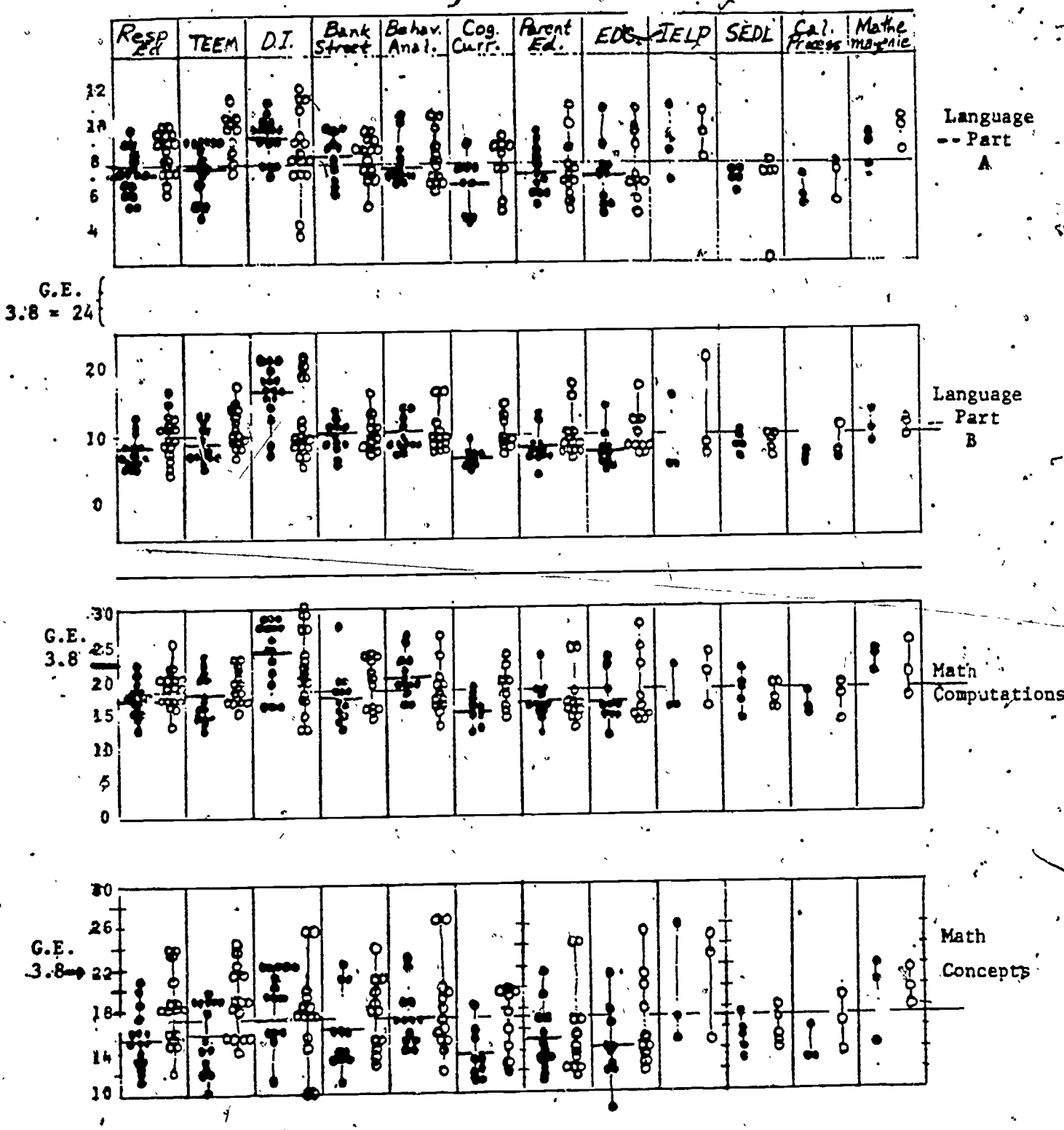

Fig. 1 Plots of site means for FT (solid circles) and NFT (open circles), data from Abt Associates Inc. IV-C, 1977, and III, 1976. Means are of raw scores on the Metropolitan Achievement Tests (Elementary version) administered at the end of third grade. Data are included from cohorts II-K, II-EF, III-K and III-EF. Horizontal lines indicate averages of site means.

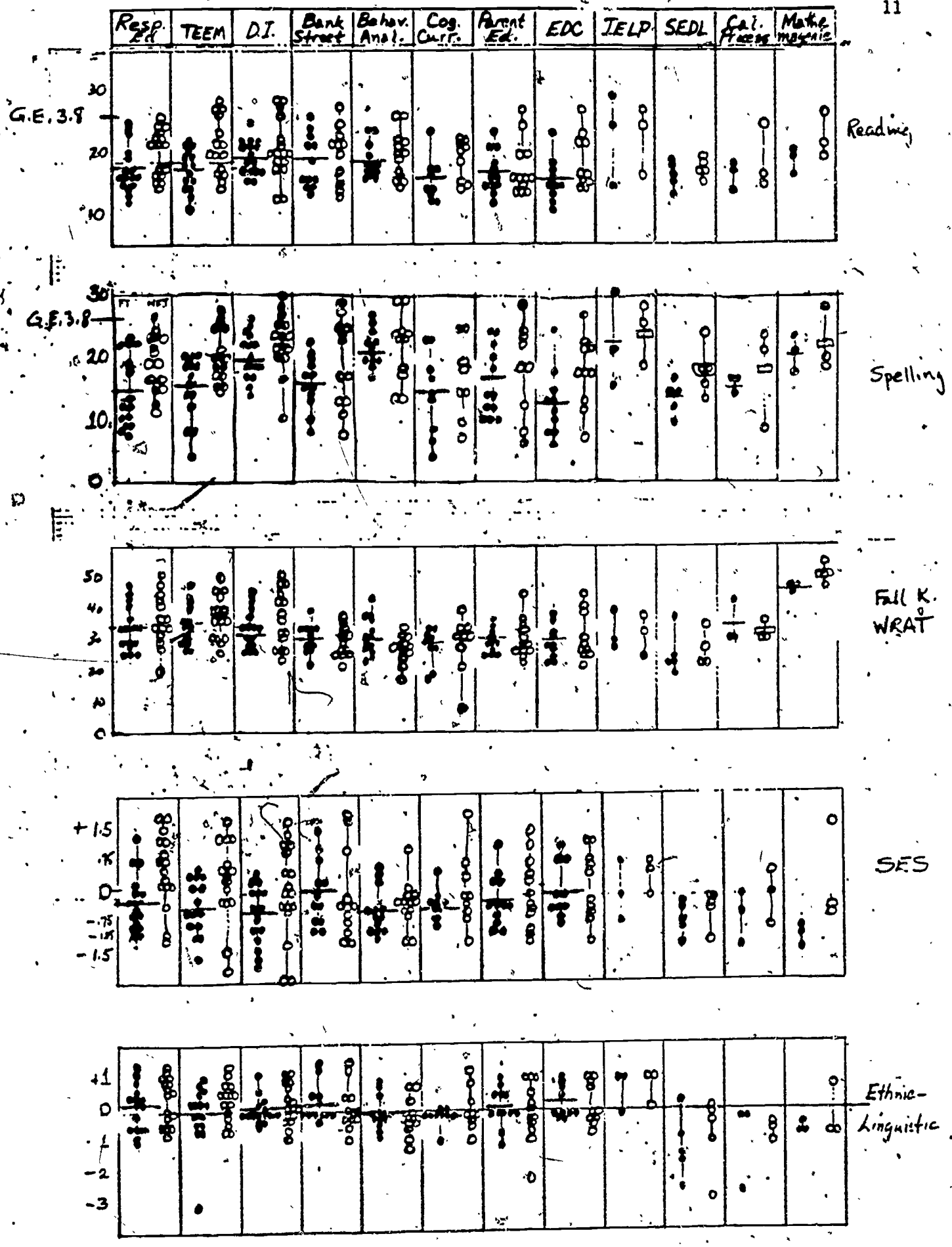

Fig. 2. Contuation of plots of site means as in Figure 1.

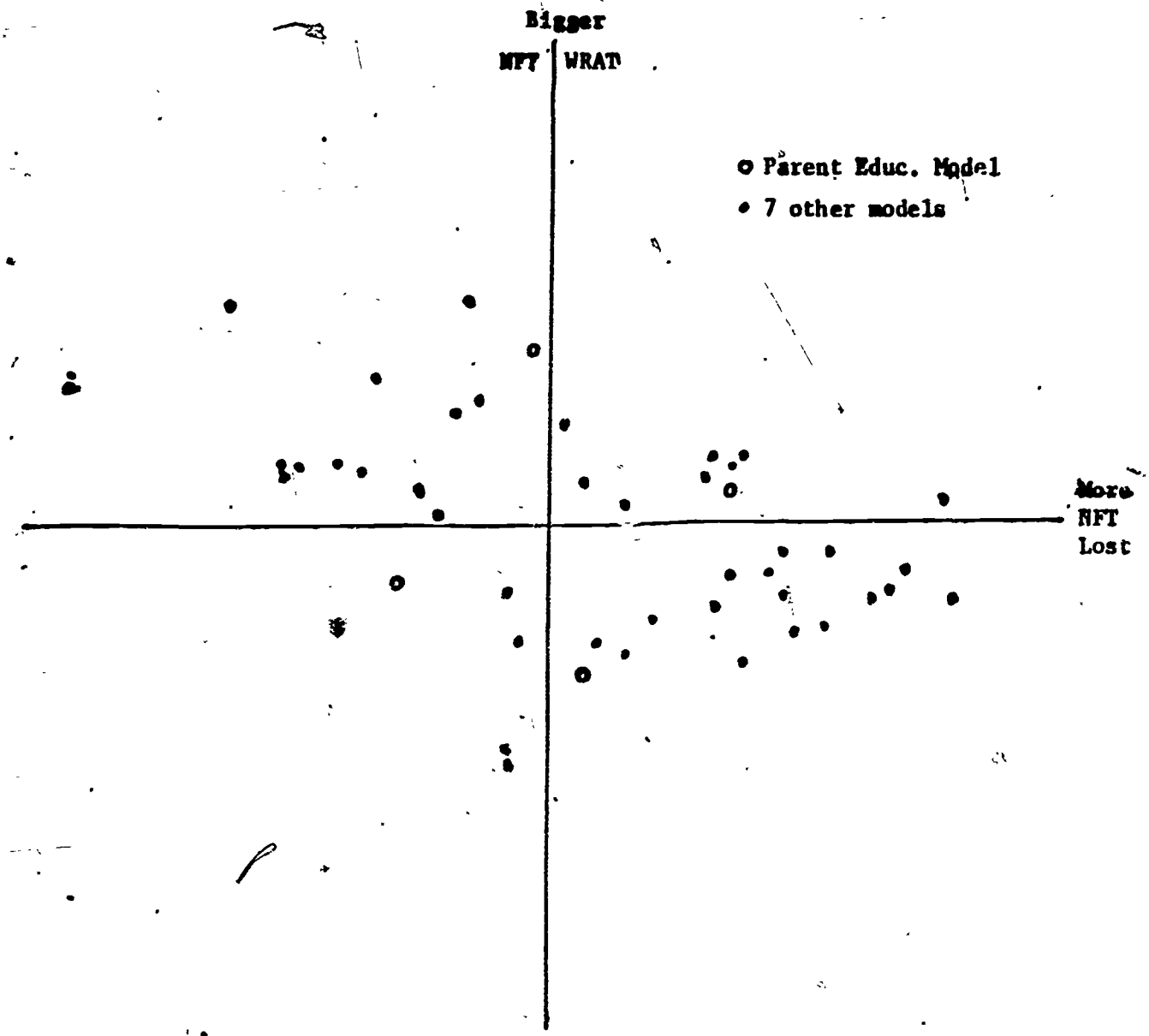

Fig. 3. Differential attrition (horizonal axis) plotted against differential in WRAT mean scores (vertical axis) at sites for eight largest Follow Through sponsors. Differential = NFT - FT. Sites where the NFT had lower WRAT scores than FT had higher NFT attrition. (r = -.51)

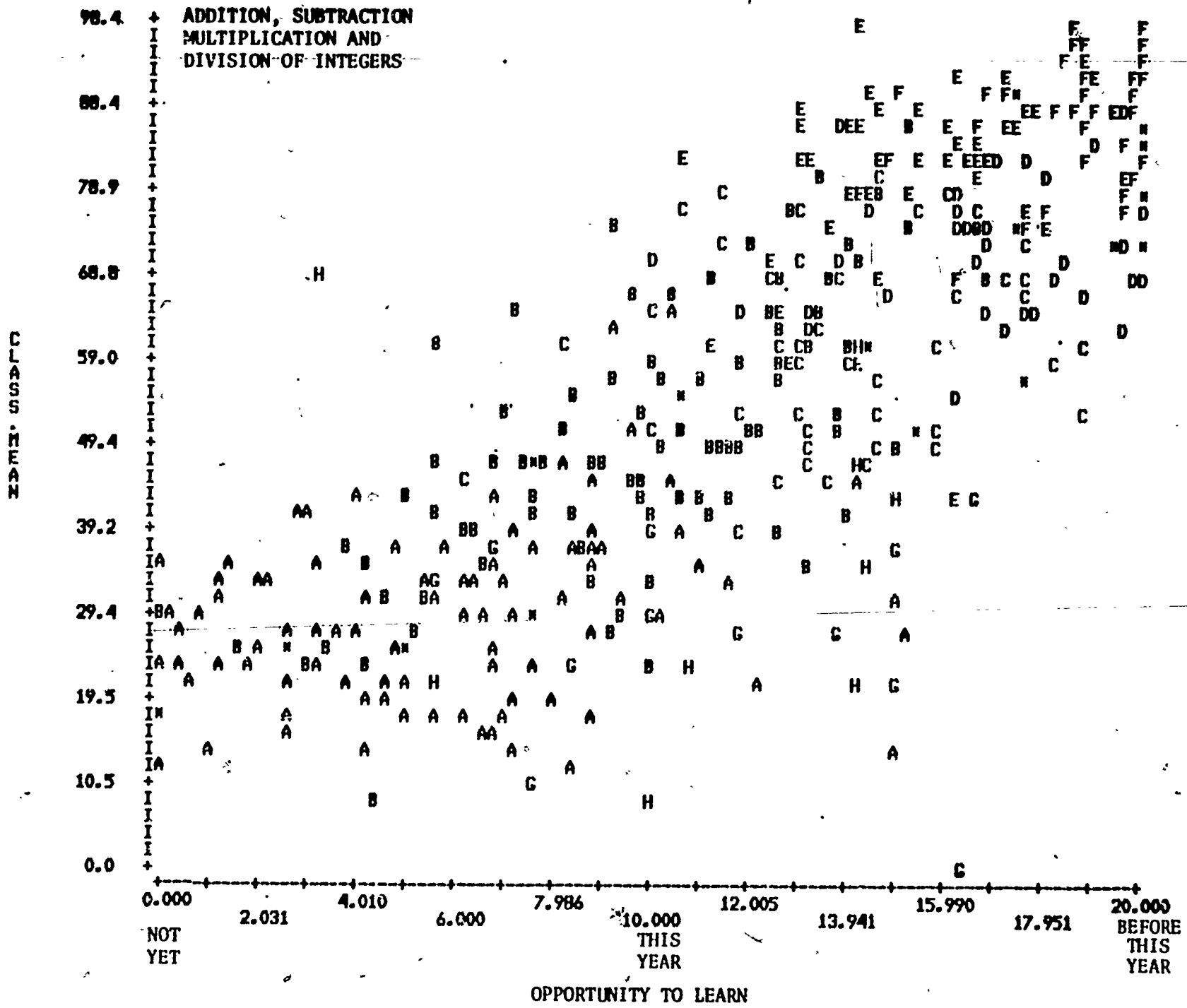

Figure 4: "Scatterplot" showing strong relationship between student reports when material was taught and student achievement.